

The University of Chicago

APPLICABILITY OF
THE PROBABLE ERROR FORMULAE
TO PSYCHOLOGICAL DATA

A DISSERTATION SUBMITTED TO THE
GRADUATE FACULTY IN CANDIDACY FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

1931

By

JAMES THOMAS RUSSELL

Private Edition, Distributed by
THE UNIVERSITY OF CHICAGO LIBRARIES
CHICAGO, ILLINOIS

1934

The University of Chicago

APPLICABILITY OF
THE PROBABLE ERROR FORMULAE
TO PSYCHOLOGICAL DATA

A DISSERTATION SUBMITTED TO THE
GRADUATE FACULTY IN CANDIDACY FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

1931

By

JAMES THOMAS RUSSELL

Private Edition, Distributed by
THE UNIVERSITY OF CHICAGO LIBRARIES
CHICAGO, ILLINOIS

1934



CHAPTER I

INTRODUCTION

One of the major problems of statistics is that of predicting from a sample the properties of a larger group from which the sample is selected. The definition of the mean of the larger group in terms of the mean and the probable error of the mean of the sample is an illustrative example.

The theory of probability upon which the probable error formulae are based was derived primarily from games of chance, such as coin or dice throwing and roulette, and from physical measurements, as in astronomy, and secondarily from statistics such as census, insurance and trade data. The primary sources fulfill fairly adequately the conditions required by the formulae.

The probability formulae thus derived and tested have been carried over and applied to all kinds of psychological data. The applicability of the probable error concept to psychological data has been discussed in a theoretical manner by a number of psychologists and statisticians. The excerpts below are illustrative of such discussions.

Brown and Thomson (3) state that any experiment the result of which depends upon a human decision has much in common with the throw of a die. Presumably the probability formulae can be used without modification.

Otis (17) writes as follows:

The chance of a pupil knowing or not knowing the answer to a question on a test is something like the chance of a coin falling heads or tails. This is the reason for the distribution of scores of a homogeneous group of pupils resembling the distribution of heads among flipped coins. There are hundreds of causes or influences operating to determine the score of a subject. The score is affected by all previous instruction, factors making up his innate capacity to learn, interest in the subject, quality of instruction, physical condition and mood. These factors are of such nature that the presence of any one of them in the subject of a group or its absence in another is largely a matter of chance. Hence there appears to be a marked similarity between the nature of causes governing the distribution of scores of a group of pupils in a test and in a set of coins repeatedly flipped.

Rugg (22) writes as follows:

Human traits are 'combinations' and include many 'arrangements' of many separate causes which may be assumed to be independent of each other. It is a safe assumption that the many separate causes (as of arithmetic ability, which is a complex resultant of causes due to hereditary capacity, physical conditions of growth and conditions of training) are independent of each other. At least they are not related in any definite way. We cannot isolate the specific unit causes but we can measure the effect of the combination of the vast number of separate causes. Human events thus are assumed to be 'compounds' analogous in their determination to compound 'chance' events of an ideal nature. The two instances show distributions of somewhat similar shape.

The above assertions may be criticized on two grounds. The first objection is that it may be as reasonable to suppose on logical or intuitive grounds that the conditions underlying routine statistical formulae are not met strictly in the case of data of the kind considered. The second objection is that distributions approximating normality may be possible in the case of data which are derived from conditions quite different from those upon which the normal curve is based. The probable

error formulae require normality of distribution of the statistics whose probable limits are specified by the formulae, hence the relevancy of the discussions of the form of distribution of the character measured. The excerpts below illustrate the second objection.

Jones (12) writes as follows:

There is no reason why the normal curve itself to which the theory (of probability) led should be subjected to precisely the same stringent limitations. After all, the real and only justification for choosing one curve rather than another to fit any given set of observations is that it does succeed in fitting them better. But when the further question is asked why the normal curve should succeed in describing some results so well we must not be tempted by analogy to rush to the conclusion that the causes at work are necessarily independent and equal, etc. In short, the theoretical justification and the empirical use of the normal curve are two quite different matters. Experience shows that the normal curve fits certain types of distributions besides those which arise in tossing coins, etc., very closely; for example, (1) certain biological statistics, (2) certain anthropometrical statistics, (3) errors of observation in experimental work, (4) marks of shots upon a target, assuming that the shots are equally liable to err in any given direction, (5) certain sociological statistics of a comparatively stationary character, (6) any statistics or quantities that are individually compounded of a large number of elements mostly independent of one another, which themselves vary between limits not very widely divergent and none of which exerts a preponderating influence upon its resulting statistic. It would be difficult in most of these cases to satisfy ourselves of the fulfillment or non-fulfillment of conditions like those upon which the binomial distribution rests. The question arises whether the above types could not, if we knew more about them, really all be included under (6), representing a further development from the binomial theory.

Thorndike (26) writes as follows:

Interdependence of the causes determining the quantity of a trait may cause all sorts of departures from normality, skewness and multimodal conditions. If

the causes are numerous and their interdependencies of a random character, their combined action may be practically identical with that of totally independent causes.

The modification of the stringent conditions underlying the derivation of the normal curve implied in these excerpts would seem to lead to results approximating in more or less degree the results which follow the stricter conditions. It is quite conceivable that a complex of an indefinitely large number of "causes", interrelated in all possible ways and degrees, would result in a normal distribution of quantities resulting from such causes. Of course, such a complex would again represent ideal conditions.

It may be argued that the atomistic conception of causes and the conception of atomistic "errors" implicit in the assumptions leading up to the formulae are questionable especially when applied to psychological, biological, etc., statistics, where it may often be reasonable to expect such characteristics as non-independence of events, inequality of causes, possible limitation of the number of adequate causes, and partial failure of the assumption that the smaller the deviation the greater the probability of its occurrence.

The two general questions which constitute the problem of this study concern the utilization of the probability concept in the case of psychological data. They are the following.

(A) Are the prevalent customs and usages of psychologists in applying the usual probable error formulae to their

data justifiable? Do these usages conform to the principles and assumptions as they have been developed and tested in reference to games of chance, etc., or do psychologists use these formulae in an illegitimate manner?

(B) Are these formulae and assumptions applicable to psychological data and problems? Can they be utilized at all, or is their utilization limited by a number of conditions? What are these limiting conditions? Are the formulae equally applicable to all psychological materials, or do the limiting conditions vary with each type of psychological material? In what respects must these formulae be modified to meet these limitations?

This paper attempts to throw some light upon the two sets of questions raised in the following manner.

(A) By means of a questionnaire we attempted to find out the customary practice of psychologists in dealing with several hypothetical psychological problems which are somewhat typical and representative of the actual problems in which the statistical concepts are employed. Is there any agreement among psychologists as to their usages? If there is any tendency to uniformity of usage, does this usage of the majority conform to the assumptions and principles of statistical theory, or if not, in what respects does it deviate from such principles? Or, what additional assumptions are made by psychologists in their application of these principles, and to what extent are these additional assumptions valid and legitimate

and to what extent are they questionable?

(B) The second question was investigated by securing a series of data from games of chance and a comparable series of various types of psychological data. These two types of material were treated in a similar statistical manner in the effort to ascertain empirically whether the two kinds of material are equally amenable to the same kind of statistical principles.

CHAPTER II

QUESTIONNAIRE

The questionnaire was submitted to psychologists in various universities. The hypothetical experiments chosen deal with memory. Part 1 concerns the probable error (PE) of the mean and Part 2 concerns the PE of the difference between means.

Thirty-three psychologists and graduate students filled out the questionnaire. The results are presented below. In the case of each question the following data are given: The number of subjects who answered the question affirmatively, the number who answered the question negatively, the number who did not answer the question, and the number of subjects whose answers were not in "yes" or "no" form but which were qualified in various ways. It is obvious that many of the questions cannot well be given categorical answers. Immediately below are given the directions which accompanied the questionnaire.

Directions: We are trying to get a representative sample of current opinion and practice concerning certain statistical concepts and operations. Read all questions of each part of the accompanying material before answering any single question. Then indicate the answer to each question on the attached sheet. Answer the questions on the basis of your own opinion as to what is true, or on the basis of what you regard

as current practice among psychologists. (The attached sheet contained the words "yes" and "no" or "true" and "false" (the case might be, for the various questions. The correct word was to be underlined.)

Part 1. The following memory experiment was performed. From the list of 2000 registrants in the College of Science, Arts and Literature at the University of Chicago during the Spring Quarter of 1931, 100 were selected by chance and tested individually at various hours during the day over a period of 2 weeks, in the same room for the same memory materials, and under the same standardized experimental conditions. The number of trials required to memorize the materials was obtained for each subject and from the 100 records the mean score of the group and its PE (PE of the distribution divided by the square root of the number of subjects) were computed. On the basis of these values, the mean and its PE, which of the following conditions are essential to a statistically valid prediction of the mean score of some second group of subjects that are likewise selected by chance? What conditions must be fulfilled in order to define statistically the mean score of the second group in terms of the mean and the PE of the mean of the first group?

1a. Is the number of subjects in the first group sufficiently large?

Yes: 27 No: 5 Blank:1

1b. What is the minimal size requisite to such a pre-

diction?

The average size given was 114; the range of answers was from 25 to 500. There were 1 blank and 4 qualified answers. 100 subjects is generally thought to be large enough for purposes of prediction.

2. Is normality of distribution of the scores of the first group essential to such a prediction?

Yes: 16 No: 17

3a. Can the second group be selected from the remaining 1900 registrants in the College of Arts, Science and Literature at the University of Chicago?

Yes: 30 No: 2 Qualified: 1

3b. Can they be selected from the total list of 2000 registrants?

Yes: 23 No: 10

3c. Can they be selected from the registrants for any quarter of the following year?

Yes: 18 No: 11 Qualified: 4

3d. Within the following 5 years?

Yes: 11 No: 19 Qualified: 3

3e. Within the following 10 years?

Yes: 11 No: 19 Qualified: 3

4a. Can the second group be selected from such Spring registrants from any American university?

Yes: 7 No: 25 Qualified: 1

4b. Only from co-educational institutions?

Yes: 7 No: 24 Qualified: 1 Blank: 1

4c. Only from co-educational institutions in the Middle West?

Yes: 4 No: 27 Qualified: 1 Blank: 1

5. Must they be tested by the same experimenter who tested the first group?

Yes: 5 No: 28

6. Must the memory materials and the experimental conditions be the same as in the case of the first group?

Yes: 33

7a. Must the size of the second group be equal to the size of the first group?

Yes: 4 No: 28 Qualified: 1

7b. Can it be larger?

Yes: 30 No: 2 Qualified: 1

7c. Can it be less?

Yes: 25 No: 7 Qualified: 1

7d. What is the minimal size that is essential to a valid statistical prediction?

The average size given was 102; the range of answers was from 25 to 300. There were 2 blank and 3 qualified answers. 100 subjects are generally thought to be enough. In the case of the individual answers, the answers to 7d corresponded to the answers to 1b except for 3 instances. In these 3 cases the size of group indicated in 7d was larger than the size of group indicated in question 1b.

8. Must the scores of the second group be normally distributed?

Yes: 17 No: 15 Qualified: 1

9a. On the basis of the mean score and its PE of any group of subjects in a memory experiment, is it ever possible to make a statistically valid generalization as to the mean score of any group of different individuals?

Yes: 26 No: 6 Qualified: 1

9b. If so, what statement can be made?

Ten subjects reported that the chances are even that the mean of the second group will lie within ± 1 PE of the mean of the first group; 6 reported that the mean of the second group will, with a certain probability, fall within certain limits. Other answers: Yes, given similar sampling and a similar number of subjects; Yes, if the 2 groups are representative of the same universe; Yes, if group B exhibits, besides normality, the same general characteristics as Group A; Yes, if there is a high degree of homogeneity of factors such as age, sex, education, general intelligence, social group; The true mean should be found in 50% of the cases (?) within ± 1 PE of the mean of group A if A equals 300 or more and B is infinite; If the PE of the distribution is assumed to be the same as that of the universe, prediction may be made to the mean of the universe; If the 2 groups are samples of the same larger population, one may say with practical certainty that the mean of the second group will fall within the range of the first group's

mean ± 6 times its own PE. The chances (?) are that it may sometime be possible.

This summary is remarkable in the range of opinion shown.

9c. Can one also make a valid predictive statement as to the size of the PE of the mean of the second group?

Yes: 24 No: 6 Blank: 2 Qualified: 1

9d. What statement can be made?

Twelve subjects indicated that their answer was the same as that to question 9b. 6 subjects stated that one can find the PE of the PE of the mean and hence determine within what limits the PE of the mean of the second group shall fall. 2 subjects stated that the PE of the mean of the second group should be about the same as the PE of the mean of the first group. 2 subjects stated that the PE of the mean of the second group should be about the same as the PE of the mean of the first group provided the 2 groups be of the same size. Other answers: The PE of the mean ought to be larger than the original PE of the mean; If the 2 groups can be considered samples of the same larger population one can say with practical certainty that the PE of the second mean will fall within the range PE of the first mean ± 6 times the PE of the first PE; Under similar conditions the most probable PE of the mean can be roughly predicted and one can calculate the PE of the PE of the mean for the second group.

Again, the answers are remarkably varied in character.

Part 2. Two groups, A and B, were selected by chance from a large population. Each group contained 100 subjects. The 2 groups served in a memorizing experiment, group A being subjected to distraction and group B being free of such distraction; but all other conditions were the same for the 2 groups. The mean of group A was 7 and the mean of group B was 9. The difference between the means was, therefore, 2 and the PE of this difference was 2. The PE of the difference was computed by the formula: PE of the difference equals the square root of the sum of the squares of the PE's of the 2 means. From this same population there was selected by chance a series of pairs of groups, X_1 and Y_1 , X_2 and Y_2 , etc. The X groups were subjected to the memory test under exactly the same conditions as was group A, and the Y groups were subjected to the test under exactly the same conditions as was group B.

1a. Are groups A and B sufficiently large to allow of a statistically valid prediction of the difference between the means of the X and Y pairs of groups?

Yes: 25 No: 7 Blank: 1

1b. What is the minimal size of the A and B groups requisite to such a prediction?

The average size indicated was 173; the range of size was from 6 to 2025. There were 2 blanks and 3 qualified answers. Comparison of the individual answers to questions 1b of Part 1 and 1b of Part 2 shows that the answers to the latter question correspond to the answers to the former question ex-

cept for 8 instances. In 5 of these 8 cases the size of group indicated in question 1b of Part 2 was larger than in question 1b of Part 1, and in 3 cases it was smaller.

2a. Must the X and Y groups each contain the same number of subjects as the A and B groups?

Yes: 3 No: 29 Qualified: 1

2b. May they contain more?

Yes: 27 No: 5 Qualified: 1

2c. May they contain less?

Yes: 20 No: 12 Qualified: 1

The answers to 2a, 2b and 2c agree in the proportion of affirmative answers to the corresponding questions in Part 1.

2d. Must groups X and Y be equal in size?

Yes: 20 No: 12 Qualified: 1

2e. What is the minimal size of the X and Y groups that is essential to a valid statistical prediction?

The average size indicated was 175; the range of size was from 25 to 2025. Four answers were qualified and 2 were blank. Comparison of the individual answers to questions 1b and 2e of Part 2 indicates that the answers to 2e corresponded to the answers to 1b except for 6 instances. In 3 of these 6 cases the size of group indicated in question 2e was larger than that indicated in question 1b, and in 3 cases it was smaller.

In the case of question 7d of Part 1 and question 2e

of Part 2, the answers to the latter question corresponded to the answers to the former question except for 5 instances. In 3 of these 5 cases the size of group indicated in question 2e of Part 2 was larger than in question 7d of Part 1, and in 2 cases it was smaller.

In the case of question 1b of Part 1, 23 of the 28 subjects who gave numerical answers indicated 100 or less as the minimum size for group A. The corresponding ratios for question 7d of Part 1, 1b and 2e of Part 2 were 23 out of 28, 22 out of 28, and 21 out of 27, respectively.

3. Given the values indicated in the experiment for the A and B groups and assuming ideal conditions, score each of the following statements as true or false.

(a) The difference between the means of the X and & groups will also be 2 in 50% of the cases.

True: 8 False: 25

(b) The X means will be smaller than the Y means in 50% of the cases.

True: 12 False: 20 Blank: 1

(c) The size of the difference will vary between 0 and 4 in 50% of the cases.

True: 27 False: 6

4a. Can one ever make a valid predictive statement as to the size of the PE of the difference of the X and Y groups?

Yes: 27 No: 4 Blank: 2

4b. What statement can be made?

Six subjects indicated that their answers to question 9b of Part 1 would also serve here. Two subjects stated that a PE of the difference similar to that of the A and B groups would be found for the X and Y groups, provided that the X and Y groups were of the same size as the A and B groups. Two subjects stated that one can determine the PE of the PE of the difference and then can approximate within what limits the PE of the difference of subsequent groups will fall. Other statements: Given the PE of the PE of the difference, the chances are even that the PE of the difference of X and Y will be greater than the PE of the difference of A and B; The PE of the difference will be within certain limits; The size of the PE of the difference of the X and Y groups should be about the same as that of the original A and B groups; The chances are even that the PE of the difference between the X and Y groups will be 2; The PE of the difference of the X and Y groups ought to be larger than the PE of the difference of the original A and B groups; Knowing the results for groups A and B alone and assuming all X and Y groups to be random samples of the same population as A and B, one may say with practical certainty that the PE of the difference between any X and any Y group will fall within the range PE of the difference of the A and B groups ± 6 times the PE of this PE.

CHAPTER III

DISCUSSION OF QUESTIONNAIRE

We shall now discuss the majority opinion on the questionnaire in the light of the statistical principles involved. There are two general questions involved in each part of the questionnaire: (1) the characteristics of the group or groups to which prediction is made on the basis of the sample; (2) the nature of the sample or samples from which prediction is made. There are thus four questions which we shall discuss in order.

1. To what group may one predict on the basis of the mean and PE of the mean of a sample? With a given sample one can predict only to the mean of the total group from which the sample was selected and which includes that sample. One can not predict from a sample to the mean of any other sample drawn from the total group by means of the customary formula.

In the case of our hypothetical example of Part 1, one can predict, on the basis of the mean and its PE, only to the mean of the 2000 registrants from which the sample was selected, and it is to be noted that this group contains the 100 individuals who compose the sample.

All questions of the questionnaire concerning the composition of the second group refer to a selected group of different individuals. It is obvious that all affirmative answers to such questions are wrong.

It is evident that many psychological data are subject to a number of serious limitations. We shall discuss two of these—the space and time limitations.

(a) Space limitation. With most psychological material, one is not justified in using a small group drawn from one locality, the University of Chicago for example, as a sample of the population of a larger area, the registrants of all colleges in the Middle West for example. In order to predict the mean score of all registrants in the Middle West, one should select a sample drawn from that total area, and not from one restricted locality of that area.

The reason is obvious. We chose data—memory scores—which presumably vary from institution to institution, and hence the scores taken from one institution can not be regarded as a legitimate sample of the scores that would be obtained from all institutions.

It is entirely possible that some characteristics with which psychologists deal may not vary with the institution or the locality but yet they may. Presumably most characteristics of an academic nature are likely to vary with the institution, and it is these with which psychologists are primarily concerned. But the fact remains that one is never justified in ignoring this spatial or institutional limitation, unless he can conclusively demonstrate that the characteristic with which he is dealing does not vary with the institution or the locality. This fact cannot be assumed; it must be experimentally proved.

The opinions as given in the questionnaire are somewhat divided in respect to this limitation, although the majority opinion is correct.

(b) Time limitation. Another serious limitation of psychological materials is that of time.

One might assume that the memory scores of our group of 100 might be taken as an adequate sample of all possible memory scores that might hereafter be made by that same group. But subsequent scores will vary with the materials used, with practice, and especially with the later stages of life. The given scores can not be regarded as such a sample.

Neither can these memory scores be taken as a sample of the mean score of Chicago registrants during the period of the next 10 years. These registrants represent a selected group, and an institution may change its standards from time to time. Likewise the character of the population seeking entrance to an institution may vary from time to time. This constitutes an insuperable limitation, for obviously one can not obtain an actual sample of future events.

It is noteworthy that the majority opinion of the questionnaire exhibited some scepticism of temporal predictions, and that the degree of scepticism increased with the length of the time interval.

Psychological material is also subject to many similar types of limitation, such as age, race, cultural status, etc. Obviously one can not take the records of our 100 subjects as

representative of the scores of young children, Indians, or a group of day laborers.

Most psychological materials thus differ from those of games of chance in one very important respect. In psychology we are dealing with samples of a very restricted group or population, while in games of chance we may deal with an unrestricted or infinite group or universe. A sample of throws of a set of dice are representative of all possible throws irrespective of the time or place or person who makes the throw.

The probable error formula given in the questionnaire can properly be utilized only in predicting from a sample to a universe—an unrestricted or infinite group. It can not be employed legitimately in predicting to a small restricted or finite group. One can not predict the mean of the 2000 registrants by such a formula. The formula must be modified in order to predict to a limited group as in the case of most psychological materials. In this case the formula for the PE_m reads: $PE_m = .6745 \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$, where standard deviation is the standard deviation of the measure of the sample, n is the size of the sample and N is the size of the population. Appropriate modifications should be made for other statistical constants. Psychologists for the most part have ignored the fact that they are dealing with samples of relatively small restricted groups, and have erred in using the formula adapted to the statistics of games of chance.

2. The characteristics of the sample.

The probable error formula enables one to predict the size of the mean of the total group relative to the mean and its PE of a sample drawn from that group. We predict that the chances are even that the mean of the total group will fall within the range of ± 1 PE of the mean of the sample, or that the mean of the total group will fall within the range of ± 1 PE of the means of all possible samples that may be drawn from that total group.

For such a prediction to be mathematically exact, the formulae require that the variability of the measures of the sample shall be exactly equal to the variability of the population from which the sample is drawn. But the standard deviation of the sample only approximates that of the total population, and hence the prediction is only approximately or relatively accurate. In other words, all predictions made on the basis of the mean and the PE of the mean of a sample are subject to PE's of their own.

The degree to which the two standard deviations approximate each other varies with the size of the sample, all other factors being constant. Also, the degree of accuracy or exactness of prediction varies with this degree of approximation. It follows, therefore, that the degree of accuracy of the above prediction will vary with the size of the sample. It is thus obvious that one can specify the requisite size of a sample only in relation to some given degree of accuracy of prediction that may be desired.

The majority of subjects were of the opinion that samples of approximately 100 cases are sufficient for purposes of prediction.

Many psychological experiments are undertaken with groups of 15 to 30 subjects and the PE formulae are often employed in these instances. It should be noted that, in the case of small samples, a modification of the usual formulae should be employed. The modification enters into the calculation of the standard deviation of the sample. The customary formula for the standard deviation of a sample is: $\sigma = \sqrt{\frac{Sd^2}{n}}$, where Sd^2 is the sum of the squares of the deviations of the measures from the mean and n is the size of the sample. In the case of small samples, the formula is: $\sigma = \sqrt{\frac{Sd^2}{n-1}}$. The former formula is, in any case, an approximation to this formula.

It is dangerous to apply the PE formulae to cases of small samples where little can be known as regards the form of distribution, with the conclusion that a certain probability applies to a particular statistic of the population.

The PE formulae apply only in the case of normal distributions of the statistic whose PE is concerned. Hence, all of the conditions underlying the normal law of simple sampling are assumed. The fundamental assumptions concerning the distributions of accidental errors, leading to the normal law of error, are variously stated and are based upon the results of experiments and upon theoretical considerations. The term error is used in the sense that if we take the mean of a number of

observations the deviation of any one of them from the mean may be termed its error.

The assumptions which have been stated are as follows:

(1) the causes of errors of measurement are independent of each other; (2) there is a very large number of causes; (3) the arithmetic mean is the most probable value of the quantity being measured; (4) errors of the same magnitude are equally likely, i.e., positive and negative errors of the same magnitude are equally likely, thus making the distribution symmetrical; (5) the smaller the error the greater the likelihood of its occurrence.

The condition of sampling required by the PE formulae is that the sample shall be adequately representative of the universe. The representative sample is composed of an adequately large number of members of the universe chosen at random with respect to the character measured.

It is often assumed that a random selection means that a group is selected without plan. This assumption is erroneous. For example, we could have arranged our 2000 registrants in alphabetical order and chosen every twentieth one. This would have been selection according to a definite plan. But it is obviously proper only if the method of selection is not correlated with the characteristics studied. Random choice implies lack of correlation between the process of selection and the character measured. The plan of selection of the sample should be specified in any case, in order that the possibility of

correlation between the plan and the characteristic studied may be investigated if necessary.

Yule (31) states the following as the conditions of simple sampling: (1) complete independence of the events composing the sample; (2) there should be no essential difference between the localities from which the members of the sample are taken, and there should be no essential change in underlying conditions during the period of time covered by the observations; (3) the conditions that regulate the appearance of the character observed are the same for every sample and also for every individual in every sample.

Yule writes as follows concerning the condition of independence:

The distribution of means drawn under the conditions of simple sampling will always be more symmetrical than the distribution of the character measured and the symmetry increases with the number of cases in the sample. The distribution of means, and therefore also of the differences between means, tends to become not merely symmetrical but normal. The approach to normality is rapid only if the several drawings for each sample are independent. If the original distribution is normal, the distribution of means, even of small samples, is normal. If we do not follow the conditions of simple sampling but first draw a series of samples from one record, then another series from another record with a different mean and standard deviation, and so on, or if we draw the successive samples from essentially different parts of the same record the standard deviation of the mean will be greatly increased. If we form a set of samples by taking one measure from each of several parts of the same record (the parts differing more or less) for the first sample, another measure from each part for the second sample, etc., the standard deviation of the means of the samples will be less than the standard deviation of simple sampling. In case there is correlation between the measures, for example, if the first measure in any sample is of high value, the other measures of the same sample are likely to be high values also, the formula is: $\sigma_m^2 = \frac{\sigma^2}{n} [1 + r(n-1)]$,

assuming that r is the arithmetic mean of all of the correlations involved. Positive correlation is introduced if we draw successive samples from different records, although the drawings of the several measures at each sampling are independent of one another. If each measure is drawn from a separate and distinct part of the record, the correlation between any two measures will, on the average, be negative.

If samples are drawn at random from a population, no matter how distributed, the distributions of the statistical constants of the samples rapidly approach normality as the samples increase in size. See Isserlis (11). The approach to normality of the distribution of means of samples taken from a non-normal population is retarded by correlation between the individuals composing the samples.

Experimentation is required to determine whether a certain method of selection is appropriate for a given character. One cannot safely make suppositions as to sampling in the absence of knowledge. The method of selection of the sample should be defined rigorously, otherwise there is no check upon the possibility of correlation between the plan of selection and the factor investigated.

The use of the PE formulae in prediction requires independence among the various items and among the various samplings. These conditions are difficult to meet and to test in the case of psychological materials. Psychology deals usually with very complex materials, the elements of which are related in very various ways. Such interrelations are difficult to measure and are usually neglected in the application of the formulae.

3. The above remarks are applicable to the question as to the characteristics of the samples of Part 2 and this question needs no further discussion.

4. The final question concerns the nature of the prediction that may be made on the basis of the difference between the means of two samples and the PE of that difference.

In our illustration the means of the two samples, A and B, were 7 and 9 respectively, and the PE of their difference was 2. The following predictions may be made: (1) if these two samples were increased in size so that each contained the total populations from which they were drawn, the chances are equal that the obtained difference between the means of the two populations lies between the values of 2 ± 2 , or between the limits of 0 and 4, with the advantage in favor of the B population in respect to the size of the mean; (2) it would follow that the chances are 3 to 1 that the mean of the B population is greater than that of the population from which the A sample was drawn; (3) finally, if we draw from the two total populations a large series of pairs of samples similar to A and B, we may predict that the Y means will be larger than the X means in 75% of the cases.

Prediction is made, in the case of the mean and PE of the mean of a sample, only to the mean of the population from which the sample is drawn, but in the case of the difference between means and the PE of the difference, prediction can be made also to the difference between the means of other pairs of

samples.

The results of the questionnaire show that there is considerable confusion and error in regard to the interpretation of the PE_d . One fourth of the subjects thought that the difference between the mean of the X and Y groups would remain identical with that of the A and B groups in 50% of the cases. This is more or less obviously false. One third of the subjects thought that the X means would be smaller than the Y means in 50% of the cases. This also is erroneous. As shown, the percentage in this particular case is 75. For cases in which the difference between the means and the PE_d are not equal the percentage would have to be determined from the usual tables. The third statement concerning the meaning of the PE_d is meaningless when analyzed carefully, although most of the subjects supposed it to be correct. This statement reads: The size of the difference will vary between 0 and 4 in 50% of the cases. The statement is true only for the difference between the means of the two populations.

It would seem best, in expressing the significance of the difference between means to indicate the probability that the mean of one population is greater than the mean of the other population. The meaning of this is clear. Less clear is the statement that the probability is so much that the true difference is greater than zero. This statement does not in itself indicate the direction of difference. Most obscure is the statement based upon a critical value of the ratio of the difference

to the PE_d that the difference is significant.

A rule widely used in practice is to treat a value of the ratio of the difference between two means to the PE_d of 3 or more (or of 4 or more) as indicative of a real difference between the two populations. It should be obvious that the choice of a critical value of the ratio of 3, for example, as indicative of a true difference, so far as the purpose of control or the establishment of a law is concerned, is purely arbitrary. Such a choice means neglect of 4.3%. It would seem better to simply express the probability of real difference, the data permitting. If critical values must be assigned, they should vary with the purpose to which the statistics are to be put. In view of the character of psychological material it may be unsafe to take for granted that a deviation exceeding 3 PE is practically impossible and that a deviation exceeding 4 PE is unheard of.

The PE_d between two means is a function of the PE 's of the means of the two samples. Hence, the previous discussion of the size of sample required for prediction of the mean applies in the case of prediction of the difference between means. Moreover, the same geographical and time limitations of the groups to which prediction is made and the method of selection of the groups that apply in the case of the PE_m also apply in the case of the PE_d .

However, the following distinctions should be noted. There may be a geographical error in respect to the PE_m but not

an error in respect to the PE_d between the means of two samples, and vice versa. For example, learning ability may vary with the university, but if the relationship of learning scores to temporal distribution of trials is investigated, the latter factor may exert the same effect upon learning scores in all universities. The mean height of one geographical group cannot be taken as a sample of the height throughout the country. But, if the relationship between height and sex is studied, the difference between the means of the two sex groups drawn from one geographical section may be representative of the entire population. The difference between intelligence rating or scholarship of athletes and non-athletes may presumably vary with the university. The memorizing ability of the students of two universities may be practically the same, and yet the differential ability to learn through the eye and the ear may vary with the university according to a difference in their methods of instruction.

CHAPTER IV

EMPIRICAL DATA

The empirical data were secured to throw light upon the second general question raised in the Introduction. The following issues are involved: To what extent are the PE formulae and assumptions applicable to psychological data and problems? To what extent are there limiting conditions? What are the limiting conditions? Are the formulae equally applicable to all psychological materials, or do the limiting conditions vary with the type of material? Psychological data and "dice" data were employed and compared.

The psychological material was so chosen that a large amount of data of varied psychological character could readily be secured. The accompanying "Test Blank and Directions" was used to procure the data, the test being given as a group test. The test consists of 27 sub-tests which represent categories that may be named broadly for descriptive purposes as chance, preference, judgment and ability. In the test as arranged below, the sub-tests are arranged together according to the category. Originally, they were distributed in random order, except that the "ability" tests were together at the end of the test. The latter tests were time limited, with the exception of sub-test 27. The categories are named to indicate in a general way the nature of the material and the fact of its miscellaneous character.

Each sub-test consists of 6 items, except that sub-test 17 contains 5 items (occasioned by printer's error), and that sub-test 27 contains 15 items. The sub-tests are arranged so that the majority have 6 possible scores to correspond to the 6 scores assignable to the 6 faces of a die. However, scores of 0 are possible on sub-tests 25 and 27, thus making the score ranges of these tests 0 to 6 and 0 to 15 respectively.

The test and the directions are given below. The directions are stated first. Then, the sub-tests are given. Originally, the directions for each sub-test and the items of the test were separate; here, they are combined.

Directions

Provide each subject with a test blank. First require that the top line be filled in: Age, Sex, Class Standing, University and Tester.

Announce the following:

The object of this test is to secure a large amount of statistical data on several types of mental reaction. The first type, for example (here specify two sub-tests), requires that you make a random selection of one item. The second type, for example (here specify a sub-test), involves preference. There is no correct answer in this case. Simply indicate the item which you prefer. If you have no genuine preference, select one item at random. The third type, for example (here specify a sub-test), involves judgment. Score the item which best represents your judgment. The fourth type, for example (here specify a sub-test), involves ability. If you do not know the answer, estimate or guess. Some of the problems under this type require calculation. They must be solved mentally. Be sure that you score one item in each problem of the test. All will work on the same problem at the same time. Listen carefully while I read the problem. Each problem will be read only once.

For example, the tester says: "Problem 1. Draw a circle around any one of the figures you choose." Then proceed to Problem 2. And so on through the first twenty problems.

The remaining problems are time limited. Upon completing the twentieth problem, announce: "Each of the remaining problems has a time limit. After reading each problem, I shall say "Start" and at the end of the time limit I shall say "Stop." No more work may be done on that problem after I say "Stop." If you do not have the answer by that time, immediately record your estimate or guess.

The Test

Chance Sub-tests

- (1) Draw a circle around any one of the numbers you choose.

14 28 19 56 31 45

- (2) Draw a line through any one of the syllables you choose.

nog loc qeg hax rel zab

- (3) Underline any one of the words you choose.

and or even but so besides

- (4) Cross out any one of the numbers you choose.

2 4 5 6 8 10

- (5) Underscore any one of the words you choose.

money is the root of evil

- (6) Place a cross in any one of the blocks of the figure.

This figure, which cannot well be shown, is a rectangle consisting of 6 compartments, 3 across and 2 deep. The scores 1 to 6 were assigned arbitrarily to these compart-

ments as follows: 1 3 5
 2 4 6

Unclassified Sub-tests

- (1) The color of red is most similar to the taste of which one of the substances? Underline the word chosen.

sugar vinegar onion bacon celery pineapple

- (2) "Carmaq" is a noun in the language of a savage people. Underline the English word which you guess to be its equivalent.

joy dog arrow woman head boat

Preference Sub-tests

- (1) Draw a circle around the color which you most prefer.

brown blue orange green red yellow

- (2) Underline the one fruit which you like best.

apple pear plum orange peach grape

Judgment Sub-tests

- (1) Underline the one possession which you believe most worthwhile.

riches fame knowledge goodness insight health

- (2) Underline the creature which you think is most intelligent.

ant cat chimpanzee elephant rat homing pigeon

- (3) Underline the word which you believe represents the best characterization of Lincoln.

candor honesty shrewdness justice humor mercy

- (4) Underline the act which you believe is the worst.

bootlegging—accepting bribes—beating a horse—picking
pockets—reckless driving—cheating

- (5) The quotation "Death tramples it to fragments" is an example of which one of the words? Underline answer.

irony hyperbole antithesis personification
synecdoche apotheosis

- (6) Underline the one word which is least like the others.

sympathy wisdom honesty frankness justice kindness

- (7) Cross out the number which you believe most closely approximates the present population of France. (The numbers represent millions.) 100 130 90 80 120 110

- (8) Estimate in inches the length of the line. Draw a circle around the number which represents your answer most closely.

_____ 2.0 2.8 2.4 2.6 3.2

- (9) Underline the word which is the closest synonym of "clever."

Dexterous smart intelligent bright keen shrewd

- (10) "Someday man may get along without sleep." What is the probability of the truth of this statement on a scale from 1 to 6, where 1 means impossible and 6 means undoubtedly true? Draw a circle around the number corresponding to your answer.

1 2 3 4 5 6

Ability Sub-tests

Note: The remaining problems are time limited. The time to be allowed is indicated in parentheses following the directions.

- (21) A cube of 5 blocks on a side is painted on the outside.

How many cubes bear paint on two and only two sides?

Cross out the number which most closely represents your answer or estimate. (90 seconds)

40 24 32 36 44 28

- (22) Imagine 6 points in space. How many straight lines can be drawn connecting the points? Draw a circle around the number representing your answer. (60 seconds)

9 6 18 21 15 12

- (23) The numbers in the series follow a certain principle.

Place a circle around the number which destroys the sequence as governed by the principle. (30 seconds)

2 4 12 46 16 8

- (24) If a man walks at the rate of $4\frac{1}{4}$ miles an hour, how many hours will it take him to walk 27 miles? Draw a circle around the number which most closely approximates your answer. (30 seconds)

5.4 5.7 5.5 6.1 5.3 5.6

- (25) Under each of the numbers write its square. (120 seconds)

13 15 17 14 16 18

- (26) The paragraph contains errors of spelling, punctuation and grammar. Add the number of errors and draw a circle around the number that best represents your answer. (90 seconds)

Between you the gate post and I, said albert, that realtor swindeld the accomodating citisens from fourty dollars. If that don't illustrate reprihensible principals Im Caesars ghost. Yes, I replied it's a lot of money he cheated them out of however everybody are happy.

19 31 23 15 27 35

(27) I shall read a list of words twice. Listen carefully, and after I finish, write in the blank space at the end of the test as many of the words as you can recall in any order. (To the tester: Enunciate the words slowly, at the rate of one every two seconds. After the first reading, say: "I shall read the words again.")

The list is as follows: harp play lion page knife
soap dust ink care tug eat talk club train tent

The test was administered to 2000 subjects under the following conditions. The test was sent to psychologists in 14 American universities and colleges and 1 Canadian university in October. One hundred records were secured from each institution, except that 200 records were obtained from one of the universities and 500 from another. The test was administered by the psychologists to their classes in general psychology. The number of administrators was usually 1 to 3 for any one institution. The conditions of administration are already described.

Scores or numerical values for the majority of the sub-tests were assigned on the basis of the position of the item chosen. For example, if a subject marked item 56 in the first sub-test, his "score" was 4; i.e., item 56 is the fourth item from the left in the sub-test. "True scores" were assigned in sub-tests Ability 5 and 7; i.e., the number of correct answers in Ability 5 sub-tests and the number of words correctly remembered in Ability 7 sub-test. Two other types of numerical

measures were used in the case of several sub-tests. In the case of sub-tests Judgment 7 and 8 and Ability 1, 2, 4 and 6 the items were rearranged so that the progression of values from left to right was in regular order and "scores" were then assigned according to position. These cases are designated by the symbol "a." In the case of Sub-tests Ability 1, 2, 4 and 6 scores were given also on the basis of accuracy of response.

Eight experiments involving chance throws or draws were undertaken. For convenience of reference these 8 experiments will all be designated as dice experiments. In two experiments, ordinary well balanced dice were used. In one case 4 dice were thrown simultaneously and repeatedly until 100 records were obtained. The procedure was repeated until 23 groups of 100 records each were obtained. This experiment is designated as "4 Dice." The scores ranged from 1 to 6 according to the values on the faces of the dice. In the other case 20 dice were thrown simultaneously and repeatedly until 23 groups of 100 records each were obtained. The designation is "20 dice." Scores were assigned in the same way as in the case of 4 dice.

Three experiments were performed with carefully cut cardboards 1 cm. square. In one case cardboards were drawn one at a time from a box containing cardboards numbered from 1 to 6. There were 50 cardboards for each number. After each draw the cardboards were thoroughly shaken. A similar experiment was carried through for cardboards numbered from 1 to 12, 50

cardboards being marked with each number or "score." The third experiment required 50 cardboards of each number from 1 to 50. Twenty three groups of 100 records each were obtained in each of these three experiments. The three experiments are designated 1-6 (50), 1-12 (50), 1-50 (50), respectively.

For the last three experiments, well made wooden cylinders 1.5 cm. in diameter and 0.5 cm. thick were used. For the first experiment, the drawing box contained 6 cylinders numbered from 1 to 6; for the second experiment it contained 12 cylinders numbered from 1 to 12; and for the third experiment it contained 50 cylinders numbered from 1 to 50. Single draws were made and the box was shaken well after each draw. In each experiment 23 groups of 100 records each were obtained. These three experiments are designated 1-6 (1), 1-12 (1), 1-50 (1), respectively.

In each dice experiment 23 groups of 100 records each were obtained in order to equal the number of records expected for the test. However, only 20 groups were available in the latter case.

In the case of the experiments with 4 dice, 20 dice, 1-6 (1) and 1-6 (50) the number of items was 6 in all cases. The expected mean of an indefinitely large group of records is therefore 3.5, in the case of a random selection of the values for all of these experiments. The obtained means of the 2300 records were 3.54, 3.53, 3.48 and 3.51 respectively. This suggests that the dice were well balanced. For the experiments

1-12 (1) and 1-12 (50) the expected mean is 6.5. The values found for the 2300 records were 6.53 and 6.52 respectively. For the experiments 1-50 (1) and 1-50 (50) the expected mean is 25.5. The values found for the 2300 records were 25.25 and 24.85 respectively.

The standard deviations of the distributions are 1.708, 3.452 and 14.431 for the ranges 1 to 6, 1 to 12 and 1 to 50 respectively, in the case of equal frequencies of all of the items, as may be expected for an indefinitely large group of records and random selection. For the experiments 4 dice, 20 dice, 1-6 (1) and 1-6 (50) the standard deviations of the 2300 records were 1.578, 1.611, 1.571 and 1.571 respectively. For the experiments 1-12 (1) and 1-12 (50) the standard deviations were 3.169 and 3.189. Again, the values are somewhat smaller than the expected value. For the experiments 1-50 (1) and 1-50 (50) the standard deviations were 14.257 and 14.729 respectively.

The empirical data which we have secured can be utilized to throw some light upon a number of questions which have been raised.

(1) As we have indicated in the discussion of the questionnaire psychological data are subject to spatial limitations, whereas games of chance are not. Thus, the various groups selected from different institutions are in reality samples of different total groups, and cannot be regarded as samples from the same population. Our empirical data support the validity of

this proposition in a number of ways.

(a) Comparison was made between the psychological test data and the dice data, with respect to the extent to which they meet the conditions of the formula of the PE of the mean (PE_m). The former groups are spatially selected but the latter are not.

The usual formula of the PE_m was used. As already noted, this formula is applicable strictly in predicting the mean of a group of unlimited size. The formula should be modified in case prediction is made to a group of restricted size. In the case of groups of 100 and a total group of 2000, the employment of the modified formula would result in a reduction of the PE_m of 2.5% approximately.

For each sub-test the number of instances in which the mean of the total group fell within $\pm 1 PE_m$ of each of the 20 groups was determined and expressed as a percentage of the total number of groups. These percentages are indicated for convenience by the letter "z." The z values were calculated also for the dice experiments.

Table I is a record of the values of z for the limits $\pm 1 PE_m$ for each sub-test. The values are indicated for groups of 100, 200 and 300 records. The groups of 200 and 300 records were obtained by combining by random choice of groups the 20 groups of 100 records each into 10 groups of 200 records each and also into 6 groups of 300 records each. Of course, it was necessary to discard 2 groups of 100 records each in making up

the groups of 300. At the foot of the table are shown the average values of z . The number of groups involved in determining each value of z also is indicated. Table IV is a corresponding record for the dice experiments.

TABLE I

Class	Sub-test	1 PE_m n=100	1 PE_m n=200	1 PE_m n=300
Chance.....	1	40	40	50
"	2	50	40	50
"	3	40	50	50
"	4	45	50	50
"	5	40	60	50
"	6	50	20	50
Preference.....	1	20	10	33
"	2	40	30	67
Judgment.....	1	45	10	33
"	2	35	20	33
"	3	25	30	33
"	4	25	60	33
"	5	30	40	0
"	6	60	50	83
"	7	25	30	33
"	7a	30	10	17
"	8	40	60	17
"	8a	45	50	17
"	9	40	40	33
"	10	45	50	50
Ability.....	1	20	20	33
"	1a	25	60	50
"	2	35	20	50
"	2a	20	40	33
"	3	50	40	67
"	4	50	60	33
"	4a	50	40	50
"	5	20	30	33
"	6	30	20	50
"	6a	25	10	17
"	7	20	10	17
Unclassified.....	1	40	50	17
"	2	20	20	17
Average.....		35.8	35.5	37.9
Number of Groups....		20	10	6

TABLE II

Class	Sub-test	1 PE _m n=100	2 PE _m n=100	3 PE _m n=100
Chance.....	1	40	85	85
"	2	50	90	100
"	3	40	80	80
"	4	45	75	90
"	5	40	85	90
"	6	50	85	100
Preference.....	1	20	50	75
"	2	40	60	85
Judgment.....	1	45	60	75
"	2	35	70	85
"	3	25	55	85
"	4	25	85	90
"	5	30	50	70
"	6	60	95	100
"	7	25	85	100
"	7a	30	50	75
"	8	40	75	90
"	8a	45	65	90
"	9	40	70	75
"	10	45	75	90
Ability.....	1	20	40	95
"	1a	25	75	85
"	2	35	70	85
"	2a	20	50	70
"	3	50	80	95
"	4	50	70	90
"	4a	50	75	90
"	5	20	45	65
"	6	30	70	90
"	6a	25	45	50
"	7	20	30	35
Unclassified.....	1	40	60	90
"	2	20	55	80
Average.....		35.8	67.0	83.3
Number of Groups...		20	20	20

TABLE III

Class	Sub-test	1 PE_m n=100	Class	Sub-test	1 PE_m n=100
Chance.....	1	60	Judgment.....	8a	20
"	2	60	"	9	20
"	3	60	"	10	60
"	4	80			
"	5	40	Ability.....	1	20
"	6	60	"	1a	40
			"	2	40
Preference	1	20	"	2a	20
"	2	60	"	3	60
			"	4	20
Judgment.....	1	40	"	4a	60
"	2	40	"	5	0
"	3	20	"	6	60
"	4	40	"	6a	40
"	5	40	"	7	0
"	6	60			
"	7	20	Unclassified	1	60
"	7a	20	"	2	0
"	8	20			
Average.....					38.2
Number of Groups.....					5

TABLE IV

Dice Experiment	1 PE_m n=100	1 PE_m n=200	1 PE_m n=300
4 dice.....	48	28	0
20 dice.....	52	55	72
1-6 (1).....	52	46	43
1-12 (1).....	48	55	57
1-50 (1).....	57	73	57
1-6 (50).....	26	55	43
1-12 (50).....	74	64	86
1-50 (50).....	48	73	100
Average.....	50.5	54.5	57.1
Number of Groups.....	23	11	7

TABLE V

Dice Experiment	1 PE_m n=100	2 PE_m n=100	3 PE_m n=100
4 dice.....	48	83	96
20 dice.....	52	83	96
1-6 (1).....	52	78	91
1-12 (1).....	48	87	96
1-50 (1).....	57	87	96
1-6 (50).....	26	65	96
1-12 (50).....	74	91	100
1-50 (50).....	48	87	100
Average.....	50.5	82.6	96.2
Number of Groups.....	23	23	23

In case the PE_m formula applies the expected value of z is 50. That is, the mean of the total group may be expected to fall within the limits $\pm 1 PE_m$ in 50% of the groups of 100 records each.

The space error is great in the case of the psychological materials. In Table I the average values of z are 35.8, 35.5 and 37.9 for groups of 100, 200 and 300 records respectively. The differences between the average values of z and 50 are 14.2, 14.5 and 12.1 respectively. In the case of the dice material where the formula may be expected to apply, the approximation of the values of z to 50 is close. In Table IV the average values of z are 50.5, 54.5 and 57.1 for groups of 100, 200 and 300 respectively. The differences between the average values of z and 50 are 0.5, 4.5 and 7.1 respectively. Because the usual formula was used instead of the formula appropriate for populations of restricted size, the expected value of z is slightly

less than 50. However, the departure of the values of z from the expected value is greater for the psychological data than for the dice data.

The values of z were calculated also for the limits $\pm 2 PE_m$ and $\pm 3 PE_m$. Table II is a record, for the psychological tests, of these values for the limits $\pm 1 PE_m$, $\pm 2 PE_m$ and $\pm 3 PE_m$ in the case of groups of 100. The average values of z are 35.8, 67.0 and 83.3 respectively. The differences between these values and the values expected in case the formula applies (50.0, 82.3 and 95.7) are 14.2, 15.3 and 12.4 respectively. There is no consistent tendency for the values of z to increase or decrease as the limits of fluctuation of the mean are extended. Table V is a corresponding record for the dice experiments. The difference between the average values of z and the expected values are negligible, being 0.5, 0.3 and 0.5 respectively for the above order of limits.

It is apparent that the conditions of the formula are well met by the dice experiments which contain no spatial limitation. But, in the case of the psychological material, groups drawn from different institutions are not samples of the same population. It may be that the total groups to which prediction may be made in this case must be restricted at least to the institution from which the sample was drawn; i.e., at least to the total bodies of students in courses in general psychology in the particular institutions. Prediction certainly cannot be made accurately to all such students by the method of sampling employed.

The 4 sub-tests, Ability 1, 2, 4 and 6, which were scored according to the amount of departure of the subjects' answers from the correct answer were not included in the major tables in the case of this method of scoring. The values of z for these tests are, for groups of 100 and for the limits $\pm 1 PE_m$, 55, 25, 60 and 20 respectively. The average value of z is 40.0.

It may be noted that, in the case of the dice experiments, there is no indication that the value of z changes in any regular manner as the range of scores increases.

(b) The spatial or psychological data were also compared with the non-spatial or dice data in respect to the ratio of the PE of the distribution of means to the PE_m of the total group.

If the data of any sub-test satisfy the conditions of the PE_m formula, then the ratio of the PE of the distribution of means of the 20 groups of 100 records each to the PE_m of the total group should equal the ratio of the square root of the number of cases in the total group to the square root of the number of cases in any sample. For groups of 100 and the total group of 2000 the latter ratio is $\sqrt{2000}/\sqrt{100}=4.47$. In the case of the dice experiments this ratio is $\sqrt{2300}/\sqrt{100}=4.80$.

Table VI is a record for each sub-test and each dice experiment of the above statistics. The first column indicates the sub-test or experiment, the second column indicates the PE_m of the total group, the third column indicates the PE of the

TABLE VI

Class	Sub-test	PE _m of the Total Group (A)	PE of the Distribution of Means (B)	Ratio B/A	Av. PE _m (C)	Ratio B/C
Chance.....	1	.0224	.141	6.29	.0987	1.43
"	2	.0232	.080	3.45	.1033	.77
"	3	.0248	.151	6.09	.1098	1.37
"	4	.0207	.109	5.27	.0917	1.18
"	5	.0280	.143	5.11	.1244	1.15
"	6	.0230	.095	4.13	.1024	.93
Preference...	1	.0199	.155	7.79	.0872	1.78
"	2	.0276	.190	6.88	.1217	1.56
Judgment.....	1	.0203	.148	7.29	.0893	1.66
"	2	.0234	.131	5.60	.1032	1.27
"	3	.0193	.115	5.96	.0847	1.36
"	4	.0264	.146	5.53	.1185	1.24
"	5	.0205	.164	8.00	.0892	1.84
"	6	.0205	.078	3.80	.0914	.85
"	7	.0208	.097	4.66	.0923	1.05
"	7a	.0286	.232	8.11	.1253	1.85
"	8	.0196	.107	5.46	.0868	1.23
"	8a	.0516	.318	6.16	.2210	1.44
"	9	.0291	.183	6.29	.1287	1.42
"	10	.0208	.116	5.58	.0919	1.27
Ability.....	1	.0203	.127	6.26	.0895	1.42
"	1a	.1071	.595	5.56	.4742	1.25
"	2	.0234	.157	6.71	.1038	1.51
"	2a	.0868	.700	8.06	.3811	1.84
"	3	.0128	.041	3.20	.0567	.73
"	4	.0186	.095	5.11	.0824	1.15
"	4a	.0477	.240	5.03	.2116	1.13
"	5	.0333	.324	9.73	.1449	2.24
"	6	.0222	.146	6.58	.0984	1.48
"	6a	.0676	.855	12.65	.2889	2.96
"	7	.0338	.483	14.29	.1427	3.38
Unclassified.	1	.0196	.105	5.36	.0869	1.21
"	2	.0238	.145	6.09	.1050	1.38
Average.....				6.43		1.46
Dice Experiment						
4 dice.....		.0238	.1127	4.74	.1127	1.00
20 dice.....		.0243	.1103	4.54	.1158	.95
1-6 (1).....		.0237	.1134	4.79	.1129	1.00
1-12 (1).....		.0478	.2201	4.60	.2283	.96
1-50 (1).....		.2005	.9384	4.68	.9564	.98
1-6 (50).....		.0237	.1295	5.46	.1127	1.15
1-12 (50).....		.0481	.1685	3.50	.2297	.73
1-50 (50).....		.2072	.8290	4.00	.9918	.84
Average.....				4.54		.95

distribution of means, and the fourth column indicates the ratio of the PE of the distribution of means to the PE_m of the total group. The ratios depart considerably, in the case of the psychological tests, from the value of 4.47, the average ratio being 6.43. The ratios for the dice experiments are in fair agreement with the expected ratio of 4.80, the average being 4.54. Again, the conditions of the formula are much better met by the non-spatial material than by the geographically selected material of the psychological test.

(c) The dice and the psychological materials were also compared with respect to the agreement between the PE of the distribution of means and the average PE_m of the various groups.

This test of the extent to which the data satisfy the conditions of the PE formula is shown in the last two columns of Table VI. The fifth column contains, for each sub-test and each dice experiment, the average PE_m of the various groups. If the above conditions are met, it is expected that the average PE_m should agree closely with the PE of the distribution of means. The last column of this table indicates the ratio of the PE of the distribution of means to the average PE_m . This ratio, in the case of most of the sub-tests, departs considerably from 1.00, the average being 1.46. The agreement is much better in the case of the dice experiments, the average being 0.95.

(d) Groups selected from the same institution were compared with groups geographically distributed, in the case of the

psychological data, in respect to the extent to which they meet the conditions of the formula of the PE_m .

The data from the institution which returned 5 groups of 100 records each were treated separately to determine the effect of the expected greater homogeneity resulting from restriction of the total group to the one institution upon the values of z . Table III records these values for the limits $\pm 1 PE_m$. The average value of z is 38.2 The value is 35.8 for the entire group of 2000 records distributed in samples assigned geographically to different institutions. The space error is very slightly reduced. It would seem that homogeneous samples may not even result from the individuals of a single institution, at least under the conditions of this experiment.

(e) Geographically distributed groups were compared with groups selected by chance from the total group of 2000 records, in the case of the psychological material, in respect to the extent to which they meet the conditions of the PE formula.

Four sub-tests were used. For each of these tests, the 2000 records were assigned by chance to 20 groups of 100 records each. Table VII compares the z values derived under these conditions with those derived under the condition of spatial selection of groups.

It is seen that the effect of the elimination of the space error is to cause the values of z to approximate the value of z (50) which may be expected in case the PE formula applied.

TABLE VII

Sub-test	z-for geograph- ically selected samples	z-for samples not geographically selected
Chance 1.....	40	50
Preference 1.....	20	50
Judgment 9.....	40	35
Ability 7.....	20	60
Average.....	30.0	48.8
Number of Groups..	20	20

(f) The degree of homogeneity of the various groups of each psychological test and of each dice experiment was determined also by calculating the percentage of cases in which the difference between means (taking all possible pairs of means) fell within ± 3 PE of the difference between the means (PE_d). The formula, $PE_d = \sqrt{PE_m^2 + PE_m^2}$, was used. This formula, as noted, is an approximation. Table VIII is a record of the percentages. In case the conditions of the formula are met, the percentages should equal 95.7. The average percentage for the psychological tests is 67 and for the dice experiments it is 82. The dice material meets the conditions of the formula more closely than does the psychological material.

It has been shown in a number of ways that the dice data, which were non-spatially selected, satisfied the conditions of the PE formula, and that the psychological data, which were selected from different institutions, did not meet the conditions of the formula. In the case of the psychological material, groups selected from different institutions are not

TABLE VIII

Class	Sub-test	Per cent of cases in which the difference between means fell within $\pm 3 \text{ P.E.}_d$.
Chance.....	1	74
"	2	98
"	3	68
"	4	74
"	5	55
"	6	84
Judgment.....	1	57
"	2	89
"	3	93
"	4	73
"	5	56
"	6	95
"	7	71
"	7a	49
"	8	71
"	8a	74
"	9	62
"	10	68
Ability.....	1	52
"	1a	66
"	2	64
"	2a	47
"	3	79
"	4	77
"	4a	70
"	5	50
"	6	69
"	6a	33
"	7	26
Average.....		67

Dice Experiment

4 dice.....	79
20 dice.....	82
1-6 (1).....	81
1-12 (1).....	81
1-50 (1).....	85
1-6 (50).....	69
1-12 (50).....	92
1-50 (50).....	90
Average.....	82

samples of a single population. Even groups selected from a single institution may not be samples of a single population. In order that groups may be such samples, it is necessary at least that the selection be non-spatial.

(2) Although prediction is made in the case of any statistic of a sample to the corresponding statistic of the universe, by means of the customary PE formulae, the majority opinion on the questionnaire favored the erroneous view that prediction can be made by the usual formulae and procedure from one sample to other samples of about the same size.

In order to demonstrate the error of such usage of the PE and to indicate the amount of error involved the following procedure was carried through in the case of the dice experiments. For each dice experiment and each group the number of means of the 22 groups exclusive of the given group, which fell within $\pm 1 PE_m$, $\pm 2 PE_m$ and $\pm 3 PE_m$ of the given group was determined. For each experiment the 23 values of this statistic were summed and the sum was expressed as a percentage of the total number of means considered.

The average percentages are 36.6, 36.6 and 36.3 for the limits $\pm 1 PE_m$ and for groups of 100, 200 and 300 records each. These values deviate greatly from 50. The erroneous assumption that 50% of the means of successive groups of equal size should fall within $\pm 1 PE_m$ of any given group is not borne out. The average percentages for groups of 100 records each and for the limits $\pm 1 PE_m$, $\pm 2 PE_m$ and $\pm 3 PE_m$ are 36.6, 66.5 and 84.6

respectively. The differences between the values obtained and the values expected in case the formula applies (50.0, 82.3, 95.7) are 13.4, 15.8 and 11.1 respectively. The amount of error is about the same for the various limits.

(3) What size of sample is adequate for prediction of the mean of a population?

It has been shown that a categorical answer cannot be given to this question because predictions have varying degrees of accuracy depending in part upon the size of the sample. However, if a precise statement of probability is required, then the standard deviation of the sample should equal the standard deviation of the total group to which prediction is made, and the value of the standard deviation should not be altered by increasing the size of the sample.

The average values of the square of the standard deviation for each sub-test and for each dice experiment were determined and compared for groups of 100, 200, 300 and 2000 records. Table IX shows this data. There is a general tendency for the square of the standard deviation to increase with increase in the size of the group, especially for the psychological material. In the case of the dice experiments, the value of the square of the standard deviation increased in 7 of 8 cases as the size of the group increased from 100 to 200; it increased in 5 of 8 cases as the size of group increased from 200 to 300; and it increased in 5 of 8 cases as the size of group increased from 300 to 2000. Expressed as percentages, these ratios are

TABLE IX

Class	Sub-test	n=100	n=200	n=300	n=2000
Chance.....	1	2.152	2.174	2.209	2.196
"	2	2.353	2.359	2.372	2.370
"	3	2.657	2.691	2.684	2.703
"	4	1.856	1.864	1.889	1.882
"	5	3.409	3.437	3.421	3.450
"	6	2.308	2.318	2.295	2.326
Preference..	1	1.664	1.724	1.723	1.737
"	2	3.262	3.287	3.277	3.342
Judgment....	1	1.769	1.793	1.826	1.817
"	2	2.361	2.372	2.374	2.399
"	3	1.605	1.615	1.605	1.634
"	4	3.097	3.129	3.133	3.051
"	5	1.784	1.802	1.828	1.843
"	6	1.841	1.844	1.837	1.851
"	7	1.881	1.887	1.846	1.901
"	7a	3.469	3.502	3.615	3.587
"	8	1.675	1.685	1.702	1.683
"	8a	10.779	10.867	10.957	11.673
"	9	3.657	3.684	3.663	3.730
"	10	1.865	1.874	1.889	1.894
Ability.....	1	1.771	1.791	1.782	1.807
"	1a	49.537	50.149	49.467	50.315
"	2	2.376	2.423	2.395	2.413
"	2a	31.975	32.702	32.580	33.073
"	3	.708	.711	.709	.715
"	4	1.501	1.509	1.543	1.521
"	4a	9.846	9.902	10.031	9.973
"	5	4.630	4.720	4.746	4.861
"	6	2.131	2.158	2.161	2.172
"	6a	18.554	19.238	19.524	20.071
"	7	4.517	4.751	4.784	5.022
Unclassified	1	1.670	1.681	1.678	1.694
"	2	2.429	2.446	2.469	2.477
Dice Experiment		n=100	n=200	n=300	n=2300
4 dice.....		2.843	2.850	2.835	2.873
20 dice.....		2.953	2.941	2.938	2.980
1-6 (1).....		2.806	2.814	2.819	2.834
1-12 (1).....		11.464	11.566	11.534	11.570
1-50 (1).....		201.330	203.380	204.300	201.330
1-6 (50).....		2.795	2.806	2.809	2.833
1-12 (50).....		11.623	11.764	11.831	11.685
1-50 (50).....		216.750	217.980	218.520	216.740

88, 63, and 63 respectively. In the case of the psychological tests, these percentages are 100, 55 and 75 respectively, for the above order of increase of size of the groups. Two of these percentages are greater for the psychological data, and the relative increase of the values of the square of the standard deviation from smaller to larger groups is, in general, greater for this kind of data.

It is obvious that the relative rate of increase of the standard deviation, as the size of the selected group increases constitutes another test of the homogeneity of the groups. Again, the dice data meet the conditions of the prediction formula better than do the psychological data.

On the assumption that the PE's of the series of means of finite samples are all of the same size and equal to the PE of the distribution of means, the chances are even that the mean of the population will fall within the limits of ± 1 PE_m of a given sample. If the PE's of the means of the samples are greater than the PE of the distribution of means, then the mean of the population will fall within ± 1 PE_m of more than 50% of the means of the samples, and, if smaller, less than 50%. The latter case is in general, true of the psychological data. References to Table VI will show this fact. In the great majority of the cases the PE of the distribution of means is greater than the average PE_m .

Table X records the values of the square of the standard deviation for groups of 100 and of 2000 for the 4 sub-tests of

which the 2000 records were assigned by chance to 20 groups. The table also contains the corresponding values from Table IX in which the samples were assigned to separate institutions. It is seen that the values of the square of the standard deviation for groups of 100 and for groups of 2000 agree more closely in the former case than in the latter.

TABLE X

Sub-test	n=100	n=100 n=2000 (from Table IX)	
Chance 1.....	2.200	2.152	2.196
Preference 1.....	1.768	1.684	1.737
Judgment 9.....	3.737	3.657	3.730
Ability 7.....	4.998	4.517	5.022

Reference to Tables I and IV will show that there is little change in the average value of z as the size of group increased from 100 to 300.

Smaller groups were also formed in order to determine whether the values of z would change materially. Twenty groups of 60 records each were made by choosing 60 records at random from each group of 100 records. One hundred groups of 20 records each were formed by choosing at random from each group of 100 records 5 groups of 20 records each. The values of z were calculated for only 5 sub-tests because of the amount of labor involved. The sub-tests were chosen to represent various kinds of data. The formula appropriate for groups of small size was used in these cases.

Table XI is a record of the values of z for each of the

5 sub-tests and for groups of 20, 60, 100, 200 and 300 records each, for the limits $\pm 1 P E_m$. The average values of z are 39.4, 38.0, 41.0, 28.0 and 50.0 for the groups named in the above order. There is no consistent tendency for the values to increase or decrease with the size of the group.

TABLE XI

Sub-test	n=20	n=60	n=100	n=200	n=300	n=20
Chance 1.....	40	40	40	40	50	50
Preference 2.....	44	45	40	30	67	53
Judgment 1.....	37	30	45	10	33	43
Judgment 6.....	51	50	60	50	83	47
Ability 7.....	25	25	20	10	17	33
Average.....	39.4	38.0	41.0	28.0	50.0	45.2
Number of Groups..	100	20	20	10	6	100

Another dice experiment was added to investigate the effect of a reduction of the size of the sample to 20. The 20 dice previously used were thrown 115 times to make up 115 groups of 20 records each. The total group was 2300. The mean of the population was 3.54. The value of z was 44. The corresponding value for the experiment with 20 dice for 23 groups of 100 records each was 52. This may indicate that the accuracy of prediction of the mean of the total group is decreased by reducing the size of the sample from 100 to 20.

One study has appeared recently which deals with the size of sample requisite to reliable prediction of the coefficient of correlation in the case of psychological material. Prak (20) determined the correlation between a number of tests

and the total scores of a series of other tests, for an entire group of 152 subjects as well as for groups of 30, 20, 15, 10 and 6 subjects. In this way it was possible to examine empirically how far such small groups still gave a reliable coefficient. He used the formula: PE of $r = .6745 \frac{1-r^2}{\sqrt{n}}$. He concluded that the formula for the PE actually holds in these cases, although the groups of 30 and 20 persons give a more favorable result than those smaller. However, at least 50 subjects is desirable.

The effect of the space error could again be tested in the case of the groups of 20 for the 5 sub-tests indicated in Table XI. The total group was restricted in each instance of 5 groups of 20 records each to the group of 100 records from which they were drawn. The resulting z values were summed over the entire list of 20 such total groups. The resulting values of z are recorded in the last column of Table XI. The average value is 45.2. The corresponding value from Table I is 39.4. In the latter case the groups were geographically selected. Thus, the expected effect of reduction of the space error is demonstrated.

(4) The character of distribution of the scores and of the means of successive groups may affect the reliability of the PE_m .

The application of the PE_m formula requires that the distribution of means of successive equal groups be normal. The forms of distribution of the means of the various dice ex-

periments were unimodal, but the forms of distribution of the means of some of the sub-tests in the case of the psychological material did not have well defined single modes. It was not thought worth while to secure measures of departure from normality in view of the relatively small number of groups.

Certain of the sub-tests were scored in such a way as to result in distributions of the scores having a single mode occurring in the middle region of the score range. Table XII is a record of the values of z for the sub-tests so treated, for groups of 100 and for the limits ± 1 PE_m . The table also contains the values of z secured in the case of the original method of scoring. There is a tendency for the z values to increase as a result of the above described method of scoring.

TABLE XII

Sub-test	z	z (Table I)
Chance 1.....	30	40
Preference 1.....	50	20
Preference 2.....	65	40
Judgment 1.....	50	45
Judgment 2.....	35	35
Judgment 4.....	40	25
Judgment 5.....	50	30
Judgment 6.....	55	60
Judgment 9.....	45	40
Unclassified 1.....	35	40
Unclassified 2.....	25	20
Average.....	43.6	35.9
Number of Groups.....	20	20

(5) Does the space error vary with the type of psychological material?

The three categories of chance, judgment and ability were compared with each other and with the dice experiments in several ways in order to throw light upon this question.

The average values of z in the case of the dice experiments, the chance sub-tests, the judgment sub-tests and the ability sub-tests were determined, for groups of 100 and for the limits $\pm 1 PE_m$. They are 50.5, 43.3, 37.1 and 31.4, respectively. It is seen that the value of 50 which may be expected in case the PE_m formula applies is approached more and more closely as the character of the material approaches that of games of chance.

The four classes of data were compared also in respect to the extent to which the conditions of the PE_d formula were met. The average percentages of cases in which the difference between means fell within $\pm 3 PE_d$ are 82, 76, 72 and 58 respectively, for the above order of classes. In case the conditions of the formula are met, the percentages should equal 95.7. Again, it is shown that the material satisfies the conditions of the formula more closely as the character of the material approaches that of games of chance.

The average ratios of the PE of the distribution of means to the PE_m of the total group, for the chance, judgment and ability tests are 5.06, 6.04 and 7.56 respectively. The ratio of 4.47 which may be expected in case the PE_m formula applies is approached more closely by the chance tests than by those which are of a more psychological character. The same

fact is also shown by the average ratios of the PE of the distribution of means to the average PE_m . These values are 1.14, 1.37 and 1.74 for the test categories of chance, judgment and ability, respectively. The ratio expected in case the conditions of the PE_m formula are met is unity.

We have shown that the psychological material is subject to a space error, whereas the dice material is not. The present data show that the space error varies with the kind of psychological data. Measurements of ability are most subject to the space error, whereas the chance tests are the least subject to the space error.

CONCLUSIONS

The purpose of this study, as already discussed, is to determine the amount of agreement among psychologists as to the range of applicability of PE formulae as applied to psychological experiments, to discuss the conditions required by the formulae, and to determine, for certain kinds of psychological data secured under certain conditions, the extent to which the data fulfill the conditions of the formulae. The plan was to undertake a limited study of several phases of the subject rather than a thorough study of a single phase.

The degree of reliability of the PE of the mean formula may be a function of the character of the material, of the method of scoring, of the form of distribution of the measures, of the kind of prediction made (whether to other samples, to the population, or to the universe), of the number of cases in the sample, of the extent to which the variability of the measures of the samples corresponds to the variability of the measures of the population, of the variability of the measures, of the presence or absence of correlation among the measures of the sample, of the degree of normality of distribution of the measures and of the means of successive samples, the degree of homogeneity of the samples. It is obvious that these factors are not always mutually exclusive in regard to their contribution to the degree of reliability.

It was found for the psychological material used that

the samples formed by drawing from various institutions could not, without considerable error, be regarded as samples of the same population. Groups formed within the same institution were slightly more homogeneous. Dice experiments which corresponded to the psychological tests in size of sample and number of items in the test were found on the whole to fulfill well the conditions of the PE formulae. Hence, the failure of the psychological data, on the average, to fulfill these conditions could not have been due to these factors but to failure to meet the conditions of simple sampling required by the formulae. The spatial and temporal limitations of psychological data in respect to sampling restrict greatly the populations to which prediction can be made.

A categorical answer cannot be given to the question as to what size of group is adequate for prediction of a certain statistic of the total group from which the sample is drawn. A certain size of sample simply permits of a certain reliability of prediction.

The psychological materials were subject to the space error. The degree of space error varied with the material. Tests of ability exhibited the greatest error, whereas tests involving a chance selection of the items forming the test were least subject to the space error.

The general results stated above are applicable to the PE of the difference between means because it is a function of the PE's of the means.

The customary formula for the PE of the mean is used properly in predicting from the mean and standard deviation of a sample the probability that the mean of the universe from which the sample is drawn lies within certain limits. Many experimenters erroneously use this formula to predict to the mean of a population of limited size. The amount of error involved in this usage has been shown. The customary formula must be modified to permit of such prediction.

The PE formulae to be used should vary with the number of cases in the sample, with the kind of prediction to be made, with the presence or absence of correlation among the measures.

The best interpretation and usage of the PE formulae would seem to be that the ideal conditions of simple sampling cannot be met precisely; hence, the expression of a probability as given by the formulae should be accompanied by a statement of the degree of error that may be involved in the probability. It may be better in many cases to take a sample and increase it until the change in variability of the measures becomes relatively small with increase in the number of cases and then to apply the formulae. Also, it may be profitable to break up the sample into smaller groups and investigate the degree of homogeneity of such groups and other characteristics which may bear upon the formulae.

Prediction cannot well be made in the case of many psychological materials from one locality to another. Homogeneous samples do not result from such a mode of selection of groups.

In the case of dice material, the total group may be unlimited in respect to size, time and place, but in the field of psychology the total group, in the practical sense, is very much restricted in all three aspects. These limitations require that the customary formulae be modified in the ways discussed, that the total group to which prediction is made should be defined carefully, and that the method of selection of samples be clearly described.

Making a prediction and making a generalization are much alike. A definite probability cannot usually be stated. A certain result is gotten if the conditions are fully and properly taken into account. However, conditions are not generally fully known, especially in the case of material of a psychological nature. We learn by experience that we never have knowledge of all of the conditions. There is always an unknown amount of error involved in prediction and in generalization, the amount depending upon the completeness of knowledge of the factors influencing the character investigated. The PE formulae are a valuable means of investigating the characteristics of a population from a sample. They are not, of course, expected to take account of the unknown conditions which may make a "sample" unrepresentative of a certain population or a "population" really consist of several populations.

It is usual for psychologists to use the customary PE formulae with no modifications and without consideration of their limitations. The use of the formulae in this way gives

a false sense of mathematical certainty to the conclusions that are made.

Acknowledgment

The writer is indebted to Professor Harvey Carr of the Department of Psychology for setting the problem and for great assistance in the development of the same.

BIBLIOGRAPHY

1. Boring, E.G. "Logic of the Normal Law of Error in Mental Measurement," American Journal of Psychology, XXXI (January, 1920).
2. Bowley, A.L. Elements of Statistics. New York: Charles Scribner's Sons, 1926.
3. Brown, W., and Thomson, G.H. Essentials of Mental Measurement, Chap. 2. London: Cambridge University Press, 1921.
4. Burgess, R.W. Introduction to the Mathematics of Statistics, Chaps. 11, 12, 13. Boston: Houghton, Mifflin, 1927.
5. Burnside, W. Theory of Probability, Chaps. 7, 8. London: Cambridge University Press, 1928.
6. Edgeworth, F.Y. "Law of Error," Encyclopedia Britannica, XXVIII (1902), 280.
7. Fisher, Arne. The Mathematical Theory of Probability. New York: Macmillan, 1922.
8. Fisher, R.A. Statistical Methods for Research Workers, Chap. 5. London: Oliver & Boyd, 1928.
9. Fullerton, G.S., and Cattell, J.M. On the Perception of Small Differences. Philadelphia: University of Pennsylvania Press, 1892.
10. Garrett, H.E. Statistics in Psychology and Education, Chaps. 2, 3. New York: Longmans, Green & Co., 1926.
11. Isserlis, L. "On the Value of a Mean as Calculated from a Sample," Journal of the Royal Statistical Society, LXXXI (1918), 75.
12. Jones, D.C. A First Course in Statistics. London: G. Bell & Sons, 1927.
13. Kelley, T.L. Statistical Method, Chaps. 4, 5, 12. New York: Macmillan, 1924.
14. Keynes, J.M. Treatise on Probability. London: Macmillan, 1927.
15. Mellor, J.W. Higher Mathematics, Chap. 9. New York: Longmans, Green & Co., 1905.

16. Mills, F.C. Statistical Methods Applied to Economics and Business, Chaps. 15, 16. New York: Henry Holt & Co., 1924.
17. Otis, A.S. Statistical Method in Educational Measurement, Chap. 6. Yonkers-on-Hudson, World Book Co., 1925.
18. Pearl, R. Introduction to Medical Biometry and Statistics, Chaps. 11, 11, 12. Philadelphia: W.B. Saunders Co., 1923.
19. Pearson, Karl. "Das Fehlergesetz und Seine Verallgemeinerungen durch Fechner und Pearson: A Rejoinder," Biometrika, IV (1904), 169.
20. Prak, J.L. "An Empirical Research on the Reliability of Correlation Coefficients," British Journal of Psychology, XXI (1931), 394.
21. Reitz, H.L. (et al). Handbook of Mathematical Statistics, Chaps. 1, 5, 7. Boston: Houghton, Mifflin, 1924.
22. Rugg, H.O. Statistical Methods Applied to Education, Chap. 7. Boston: Houghton, Mifflin, 1917.
23. Russell, Bertrand. "Heads or Tails," Atlantic Monthly, CXLVI, (August, 1930).
24. "Student." "The Probable Error of the Mean," Biometrika, VI (1908), 1.
25. "Student." "The Distribution of the Means of Samples Which are not Drawn at Random," Biometrika, VII (1909), 210.
26. Thorndike, E.L. An Introduction to the Theory of Mental and Social Measurements. New York: Teachers College, 1913.
27. Thurstone, L.L. The Fundamentals of Statistics, Chap. 21. New York: Macmillan, 1925.
28. Walker, H.M. Studies in the History of Statistical Method, Chap. 2. Baltimore: Williams & Williams Co., 1929.
29. Weld, L.D. Theory of Errors and Least Squares, Chap. 8. New York: Macmillan, 1926.
30. Whittaker, E.T., and Robinson, G. The Calculus of Observations. London: Blackie & Son, 1924.

31. Wilson, E.B. "The Statistical Significance of Experimental Data," Science, LVIII (1923), 93.
32. Yule, G.U. Introduction to the Theory of Statistics, Chaps. 13, 14, 15, 17. London: Charles Griffin & Co., 1924. See bibliographies at ends of chapters.
33. Zizek, F. Statistical Averages. New York: Henry Holt & Co., 1913.

